9/15/21: Data Representation, Distance & Similarity

Data Representation:

Records
- m-dimensional points/vectors
- Example: (name, age, balance) -> (“John”, 20, 100)

Graphs
- Nodes connected by edges
- Example: downward facing equilateral triangle can be represented by an adjacency matrix or an adjacency list
    - Each node is designated A, B, C
    - is there an edge between A and B, A and C, etc
    - turn that into a matrix (0 if no edge, 1 if yes edge)

Images
- A matrix or list of the pixels

Text
- list of words

Time Series
- list of data at specific intervals of time

Types of Learning
- supervised
    - turning a table of data into a graph
    - classification
- unsupervised
    - goal: find interesting structure in the data
    - ex. dataset: collection of articles
        - question: are these articles covering the same topics?

Distance and Similarity:

- unsupervised learning

feature space

- generate for all possible values for the set of features in our dataset

distance

- in order to uncover interesting features from our data, we need a way to compare data points

dissimilarity function

- function that takes 2 objects (data points) and returns a large value if these objects are dissimilar

special type: distance function

- d is a distance function if and only if:
    - d(i,j) = 0 iff i = j
    - d(i,j) = d(j, i)
    - d(i,j) <= d(i,k) + d(k,j)
        - if you go through a third point (k) to get from point i to point j, you only add distance to it
- makes intuitive sense

Minkowski Distance

- for x,y points in d-dimensional real space
    - has d components or attributes or features
- x = [x1,.....,xd] and y = [y1,.....yd]
- p>= 1
    - when p = 1, called euclidean distance
    - when p = 2, called manhattan distance